Received Date: 17-Jun-2015

Revised Date: 21-Sep-2015

Accepted Date: 22-Sep-2015

Article Type: Original Article

Corresponding author mail id: denisroy1@gmail.com

Hybrid 'superswarm' leads to rapid divergence and establishment of populations during a biological invasion

Denis Roy[1,§,*], Kay Lucek[1,2,¶], Ryan P. Walter[3], Ole Seehausen[1,2]

[1] Centre for Ecology, Evolution & Biogeochemistry
EAWAG Federal Institute of Aquatic Science and Technology
Seestrasse 79, 6074 Kastanienbaum
Switzerland

[2] Institute for Ecology and Evolution
University of Bern
Baltzerstrasse 6, 3012 Bern
Switzerland

[3] Department of Biological Science
California State University Fullerton
Fullerton, CA 92831
USA

[§] Current address: Department of Natural Resources and the Environment and Center for Environmental Sciences and Engineering
University of Connecticut
3107 Horsebarn Hill Road
Storrs, CT 06269-4210
USA

[¶] Current address: Arthur Willis Environment Centre

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.13405

Department of Animal and Plant Sciences
University of Sheffield
Western Bank
Sheffield, S10 2TN, UK

Keywords: Colonization, Hybridization, Rapid divergence, Stickleback, Switzerland

* Corresponding author

Running title: Rapidly formed populations by hybridization

Abstract

Understanding the genetic background of invading species can be crucial information clarifying why they become invasive. Intraspecific genetic admixture among lineages separated in the native ranges may promote the rate and extent of an invasion by substantially increasing standing genetic variation. Here we examine the genetic relationships among threespine stickleback that recently colonized Switzerland. This invasion results from several distinct genetic lineages that colonized multiple locations and have since undergone range expansions, where they coexist and admix in parts of their range. Using 17 microsatellites genotyped for 634 individuals collected from 17 Swiss and two non-Swiss European sites, we reconstruct the invasion of stickleback and investigate the potential and extent of admixture and hybridization among the colonizing lineages from a population genetic perspective. Specifically we test for an increase in standing genetic variation in populations where multiple lineages coexist. We find strong evidence of massive hybridization early on, followed by what appears to be recent increased genetic isolation and the formation of several new genetically distinguishable populations, consistent with a hybrid 'superswarm'. This massive hybridization and population formation event(s) occurred over approximately 140 years and likely fuelled the successful invasion of a diverse range of habitats. The implications are that multiple colonizations coupled with hybridization can lead to

the formation of new stable genetic populations potentially kick-starting speciation and adaptive radiation over a very short time.

Introduction

Populations introduced outside their species range may suffer severe genetic bottlenecks and founder effects reducing levels of standing genetic variation available for selection. This can substantially decrease the population's ability to establish and spread into novel environments (Lockwood *et al.* 2007; Dlugosch & Parker 2008; Prentis *et al.* 2008; Simberloff 2009). Consequently, many introduced species persist only locally or briefly after their introduction (Sakai *et al.* 2001; Lockwood *et al.* 2007). Some introduced species, meanwhile, establish viable populations and undergo range expansions despite initial decreases in genetic variation relative to their ancestral population (Lockwood *et al.* 2007; Dlugosch & Parker 2008). Less commonly, introduced species may colonize new geographic regions from multiple, yet genetically distinct sources, which can meet in secondary contact zones after initial range expansions. Within such contact zones, distinct lineages can hybridize converting between-lineage genetic variation to within-population genetic variance (Mallet 2007; Prentis *et al.* 2008; Abbott *et al.* 2013; Seehausen *et al.* 2014). This, in turn, increases standing genetic variation and reduces genetic constraints in newly formed hybrid populations, augmenting their genetic potential or adaptability (Mallet 2007; Prentis *et al.* 2008; Nolte & Tautz 2010; Abbott *et al.* 2013; Seehausen *et al.* 2014; Williams *et al.* 2014). Release from former genetic constraints may occur because allelic combinations fixed in parental lineages, expressed through their genetic variance-covariance matrices (VCVs), can be disrupted, the genetic covariance broken and the genetic variance broadened in ensuing hybrids (Buerkle *et al.* 2000; Abbott *et al.* 2013; Seehausen *et al.*

2014). Broadened genetic VCVs may better respond to directional selection than narrower ones especially when selection is applied off the main VCV trajectory (assuming loci reflect quantitative traits under selection with some heritability; Schluter 1996; Steppan *et al.* 2002; Schluter & Conte 2009; Seehausen *et al.* 2014). A direct prediction of this is that hybrid lines ought to have greater variance and reduced directionality (i.e., narrowness) in their genetic VCVs than their formative parental lineages (Mallet 2007; Prentis *et al.* 2008; Schluter & Conte 2009; Abbott *et al.* 2013; Seehausen *et al.* 2014).

An increased genetic potential in hybrid populations may facilitate subsequent colonization and establishment, and allow genetically admixed individuals to tap into novel niches within the invaded range not typically occupied by any of its ancestors (Lockwood *et al.* 2007; Yoder *et al.* 2010; Williams *et al.* 2014). For hybrids to persist, however, their distribution (in allopatry) and/or the balance between selection and gene flow (in sympatry or parapatry) should help establish reproductive isolation (Grant 1994; Buerkle *et al.* 2000; Mallet 2007; Nolte & Tautz 2010; Abbott *et al.* 2013; Seehausen *et al.* 2014). Otherwise, newly formed gene/trait combinations can be quickly eliminated or resorbed into parental lines (Grant 1994; Buerkle *et al.* 2000; Mallet 2007; Schluter & Conte 2009). The establishment of such newly adapted, reproductively isolated populations can ultimately lead to the formation of new species (Buerkle *et al.* 2000; Mallet 2007; Nolte & Tautz 2010; Abbott *et al.* 2013; Seehausen *et al.* 2014) and, under the right conditions, facilitate adaptive radiations (Seehausen 2004; Schluter & Conte 2009; Nolte & Tautz 2010; Abbott *et al.* 2013).

Despite an increasing number of both theoretical and empirical studies underscoring the importance of hybridization during biological invasions and species formation (Buerkle *et al.* 2000; Seehausen 2004; Mallet 2007; Seehausen *et al.* 2008; Abbott *et al.* 2013), the population genetic mechanisms operating from secondary contact to the emergence of new hybrid types remain vague (Nolte & Tautz 2010; Abbott *et al.* 2013 but see Buerkle *et al.* 2000). Thus, theoretical considerations notwithstanding, there is a need to identify systems appropriate for the study of the incipient stages of hybrid lineage formation and subsequent speciation (Buerkle *et al.* 2000; Nolte & Tautz 2010; Seehausen *et al.* 2014). The identification of newly formed hybrid lineages can not only provide key insights into the formation of new hybrid species, but also answer important questions related to the pace of hybrid lineage stabilization and the associated extent of genomic integration (Nolte & Tautz 2010; Abbott *et al.* 2013; Seehausen *et al.* 2014; Williams *et al.* 2014).

The threespine stickleback (*Gasterosteus aculeatus* species complex) has repeatedly colonized freshwater environments throughout its natural range from marine ancestors shortly after the last glacial retreat (~ 10 000 years ago). In many newly colonized freshwater habitats, stickleback have formed distinct ecotypes (McPhail 1984; Schluter 1993; Thompson *et al.* 1997; Kaeuffer *et al.* 2012; Ravinet *et al.* 2013) mostly through recurrent selection on standing genetic variation maintained at low frequencies in marine populations (Schluter & Conte 2009; Deagle *et al.* 2012; Jones *et al.* 2012). Many of the studied marine-to-freshwater stickleback colonizations have been associated with genetic bottlenecks, reducing genetic variation and likely, the adaptive potential within freshwater habitats (Reusch *et al.* 2001; Mäkinen *et al.* 2006; Deagle *et al.* 2012). While stickleback are common in many parts of Europe (Bertin 1925; Munzing 1963; Mäkinen *et al.*

2006), their distribution within Switzerland was restricted to the tributaries of the Rhine near Basel prior to 1870 (Lucek *et al.* 2010; Fig. 1). Following several introductions and the channelization of Swiss waterways (Heller 1870; Fatio 1882; Bertin 1925), stickleback underwent a range expansion and now occupy a wide range of different habitats throughout the country (Berner *et al.* 2010; Lucek *et al.* 2010; Lucek *et al.* 2013; Lucek *et al.* 2014b). The Swiss midlands are characterized by many large lakes linked by a vast network of streams and canals allowing gene flow among different lake systems, which enables adaptation to distinct habitats (e.g., shallow rivers and streams versus deep large lakes; Berner *et al.* 2010; Lucek *et al.* 2010; Lucek *et al.* 2013; Lucek *et al.* 2014b). A mitochondrial DNA survey of samples collected across the country revealed the colonization of Switzerland by three distant genetic stickleback lineages (five mtDNA haplotypes) from different parts of Europe (Lucek *et al.* 2010). The Lake Constance area is dominated by an eastern European lineage from the Baltic region (haplotype EU27; Mäkinen & Merilä 2008a; Fig. 1; Table S1), whereas the Lake Geneva area is dominated by a lineage typical of the Rhône (haplotypes EU09, EU10 and EU36). A third lineage dominates the Basel region, and may have been native to that small part of Switzerland (CH01; Lucek et al. 2010). Over the last 140 years, all three lineages have expanded into the Swiss midlands. In places such as lakes Neuchâtel, Biel, Wohlen, and in their drainages, populations are a mix of several mitochondrial lineages associated with elevated haplotype richness (Lucek *et al.* 2010). An amplified fragment length polymorphism (AFLP) analysis suggested considerable admixture between lineages in the Aare river region (near GIP and WOH; Fig. 1), wherein individuals display increased phenotypic variation (Lucek *et al.* 2010). However, the extent of admixture and the stability of population boundaries have not been previously assessed.

Here, we use a suite of microsatellite markers to infer genetic relationships among stickleback collected across Switzerland, substantially expanding on previous work relying on AFLPs (Lucek *et al.* 2010), by adding samples collected within zones showing coexistence of multiple mitochondrial lineages. First, we assess the population structure of stickleback in Switzerland in the context of known introductions. We next determine the sizes and connectivity among recovered populations assessing both their contemporary gene flow and that which has occurred in the coalescent. Finally, and in the context of previous work in the system, we examine the likelihood that some populations originate from the hybridization among main colonizing lineages as determined by Lucek et al. (2010). Overall, we show that hybridization can lead to the development of new populations whose connectivities are quickly reduced. These nascent populations may thus represent important initial steps by which colonization and hybridization work together to promote speciation, and potentially catalyze adaptive radiations over very short time scales.

Material and Methods

Sample collection & genotyping

Stickleback were collected from 17 different sampling sites across Switzerland, between summer 2007 and autumn 2008 (Fig. 1; Table S1). The sampling sites included lakes, streams and ponds. Two additional samples collected outside Switzerland were taken, representing populations to the North and South of the invaded range (Lucek *et al.* 2010; i.e., a Mediterranean freshwater population from Corsica and a North Sea derived freshwater population from Northern Germany; Fig. 1 Table S1). DNA was extracted from each individual, using a Qiagen BioSprint 96 robot with the Qiagen Blood Extraction kit (Qiagen, Switzerland). The genotype of 634 individuals

was assessed at 17 microsatellite loci using a CEQ 8000 (Beckman Coulter, Switzerland) following manufacturer instructions. The 17 microsatellites are located on 15 of 26 linkage groups determined by Peichel *et al.* 2001 and were amplified in each individual using five multiplexing sets (Table S2). Previous work has shown association between 7 of these markers and the quantitative traits of spine lengths, the number of lateral plates and gill rakers (Table S2). No evidence of null alleles, scoring errors or large allele dropouts was detected at any loci in any sampling site after checking all genotypes using MICRO-CHECKER 2.2.3 (van Oosterhout *et al.* 2004).

Population Genetic Structure

An iterative approach was used to get an unbiased, best estimate of the statistically supported number of distinguishable genetic clusters adhering to population genetic criteria (i.e., satisfying HWE and showing acceptable levels of linkage among loci). Population structure among all genotyped individuals was first assessed using STRUCTURAMA 2.0 (Huelsenbeck *et al.* 2011) which searches parameter space for the most likely number of genetic clusters using a Bayesian framework. Population number was set to a random variable but allowed to vary using a Dirichlet Process Prior (DPP). STRUCTURAMA searches used an unsupervised mode with DPPs set to 1-10, 12, 15, 17, and 20. Each search ran for 10 000 000 iterations run over three Markov Chain Monte Carlo (MCMC) sampling chains. The number of populations was collected every 100^{th} iteration collecting 100 000 values overall where the first 50 000 were discarded as burnin (Huelsenbeck & Andolfatto 2007; Huelsenbeck *et al.* 2011). The most likely number of genetic clusters recovered was determined either by consensus among searches or by selecting results of the search(es) with maximized marginal likelihoods. STRUCTURAMA analyses were

performed hierarchically by first using the entire dataset to get an overall assessment of the number of populations. All individuals were then assigned to a particular recovered cluster by their largest posterior probabilities assessed by STRUCTURE (see below), regardless of location and STRUCTURAMA analyses were then re-run on each cluster. This process was repeated until no further sub-division of clusters was observed or even genotype splitting of all individuals occurred (see Fig. 2). At each step of the hierarchical search, STRUCTURE 2.3.4 (Hubisz *et al.* 2009) was used to visualize recovered genetic clusters estimated from STRUCTURAMA and assess individual admixture proportions outlining their probabilities of belonging to recovered clusters. In STRUCTURE, the probability of each individual's assignment to recovered clusters was assessed through 10 permutations of the number of clusters recovered from STRUCTURAMA, with each permutation running over 1 000 000 iterations with an additional 100 000 used as burnin. STRUCTURE analyses allowed admixture and used correlated allele frequencies in the population structuring models. Results of all STRUCTURE permutations assessed for each hierarchical step were combined into a single individual-based clustering assignment probability using CLUMMP 1.1.2 (Jakobsson & Rosenberg 2007) and plotted using DISTRUCT 1.1 (ROSENBERG 2004).

Marker Neutrality

Seven of the markers used in this work have been previously linked to quantitative traits differing among other studied populations (Peichel *et al.* 2001; Mäkinen *et al.* 2008a). Consequently, there is a possibility that these same loci may also be linked to traits that vary within or differ between our recovered populations as well. Because the population structure recovered using markers under selection can differ from that determined using neutral markers,

(e.g., Jakobsdóttir *et al.* 2011; Bradbury *et al.* 2013; Roy *et al.* 2014) all loci were assessed for either balancing or diversifying selection. Markers were subjected to both the stepwise mutation and the infinite allele models (SMM and IAM, respectively) of microsatellite mutation and tested for neutrality using an F_{ST} outlier test (FDIST) as applied in LOSITAN 2.0 (Antao *et al.* 2008). The application of both models used 1 000 000 permutations to establish 95% confidence intervals and used a sample size reflecting the smallest genetic population under consideration. Selection affecting our markers was also tested using Bayescan 2.1 (Foll & Gaggiotti 2008) which applies a Bayesian framework to determine whether differentiation at a given locus is best explained by a model including a locus-specific component (evidence of selection) or one that is strictly related to population(s) (i.e., neutral). Bayescan assessments were set to collect every 100^{th} iteration over a total of 1 000 000 steps for a total of 10 000 recorded iterations. An additional 1 000 000 iterations were used as burnin. Priors for each assessment were adjusted using 20 pilot runs, each running 50 000 iterations. All three loci selection tests (FDIST-IAM/FDIST-SMM and Bayescan) were initially applied at the base of the recovered population structure hierarchy but also applied at deeper levels within it.

Population genetic indices and statistics

Linkage disequilibrium among loci (LD) and their adherence to Hardy-Weinberg expectations (HWE) was assessed in each genetic cluster recovered from the STRUCTURAMA/STRUCTURE analyses (hereafter populations) using Arlequin version 3.5.1.2 (Excoffier & Lischer 2010). LD tests used 10 000 permutations and deviations from HWE were tested using 1 000 000 MCMC iterations with 100 000 dememorization steps. Significance of both LD and HWE tests were assessed using sequential Bonferonni corrections

(Rice 1989). Arlequin was also used to estimate population-specific observed and expected heterozygosities (H_o and H_e, respectively). Population-specific allelic richness (with rarefaction; A_R) and inbreeding coefficients (F_{IS}) were estimated in FSTAT 2.9.3.2 (Goudet 1995). The number of private alleles (*Pa*) per population was also calculated (with rarefaction) using GenalEx 6.5 (Peakall & Smouse 2006). Levels of genetic differentiation among all possible population pairs was evaluated using the classic F_{ST} index (calculated as θ; Weir & Cockerham 1984) supported by 1000 bootstraps and derived from 100 000 permutations of the MCMC algorithm implemented in MSA 4.05 (Dieringer & Schlötterer 2003). The pairwise D_{Jost} index of genetic differentiation was also estimated with DEMEtics (Gerlach *et al.* 2010) using 1000 bootstrapping iterations to calculate significance. To test whether loci putatively linked to quantitative traits (see above) exhibited significantly different population genetic indices relative to unlinked ones, global locus-specific A_R, H_o, H_e, F_{IS} and F_{ST}s were compared using Wilcoxon sum rank tests. A_R, H_o, H_e, F_{IS} and *Pa* were also compared between Swiss populations (as inferred by STRUCTURAMA) exhibiting mtDNA haplotypes consistent with a single main colonizing lineage (hereafter MCL) versus those exhibiting the presence of multiple lineages (see Fig. 1, Table S1) using Welch's Two-sample *t*-tests.

Population Size and Connectivity

Contemporary effective population sizes (N_e) were estimated for each population using the linkage disequilibrium model (LDNe) based on single moment data as implemented in N_eEstimator v2 (Do *et al.* 2014). LDNe uses the weighted average level of expected random linkage disequilibrium among alleles over loci pairs within a given population to estimate its effective size (Waples & England 2011). Estimates of N_e based on linkage disequilibrium

assume selective neutrality, no physical linkage among loci and a closed but randomly mating population. Because our data could not identify differently aged individuals, and likely combined several year classes, our estimates most likely reflect something between the effective number of breeding individuals N_b and N_e (i.e., $\hat{N}_e$) within each population rather than the true population-specific N_e (Hare *et al.* 2011). These estimates may nevertheless be useful in gauging the relative size of populations (Hare *et al.* 2011; Do *et al.* 2014). $\hat{N}_e$ estimates were made using allele frequencies greater than 0.01 and 95% credible limits were established from jackknifing over all loci pairs. Contemporary gene flow among populations was assessed by BayesAss 3.0 (Wilson & Rannala 2003), which makes relatively few population based assumptions (e.g., populations are not required to be in HWE) and uses current allele frequencies both within and among populations to estimate recent migration rates among populations using a Bayesian approach. In BayesAss 3.0, 10 000 000 MCMC iterations were used as burnin and an additional 100 000 000 iterations were sampled at an interval of 1000. This procedure used mixing parameters of 0.3, 0.5 and 0.1 for allele frequencies, inbreeding coefficients and migration rates, respectively, and led to a total sample size 100 000 from which estimates were derived.

Coalescent-based Size and Connectivity

To generate time-integrated estimates of N_e that also consider historical influences among populations, including migration rates (m), we applied isolation with migration (IM) models estimating the long term N_e and m of each population in the coalescent (Hey & Nielsen 2004; Hey 2010). IM models search parameter space for the most likely estimates using a Bayesian framework assuming random mating within populations and that populations are each other's closest relatives not exchanging genes with other nonsampled populations (Hey & Nielsen 2004;

Hey 2010). We used IMa2 on a subsample of 9-35 individuals from each population combining their microsatellite genotypes with 436 bp of mitochondrial control region (CR) and 965 bp of cytochrome B (CytB) sequences determined by Lucek et al (2010). Although we recognize that our data may violate some of the IM model assumptions, previous work has shown that IM models as applied in IMa2, are generally robust to random mating violations and those involving small to moderate levels of introgression among considered taxa (Strasburg & Rieseberg 2010). IM analyses were run pairwise between populations following recommendations concerning the information (i.e., number of marker loci) needed for reliable parameter estimation in studies involving more than two populations (IMa2 manual; Hey 2010). Searches used priors determined from preliminary runs and were iterated using between 6 000 000 - 26 000 000 steps to reach stationary distributions before sampling. Once stationarity was achieved, all searches ran for an additional 10 000 000 steps, sampling every 100th step for a total of 100 000 recorded genealogies from which parameters were assessed. All analyses used 100 metropolis-coupled MCMC chains with heating terms ensuring high swap rates among them (<0.70). Long-term N_e and m were calculated from generated population-specific θ estimates using mutation rates of 1×10^{-4}, 9.6×10^{-6}, and 1.97×10^{-5} for microsatellites, CR and CytB sequences, respectively. Mutation rates used for the mtDNA fragments were devised by Mäkinen & Merilä (2008b) and are based on a relaxed molecular clock anchored on the estimated divergence time between *G. aculeatus* and *Pungitius pungitius* (ninespinned stickleback) determined from fossil evidence (Bell & Foster 1994; Mäkinen & Merilä 2008b). These mutation rates were used in previous studies implementing IM based analyses in other stickleback populations including some originating from the same lineages and using the same mtDNA fragments as those used here (Mäkinen & Merilä 2008b; Lucek *et al.* 2010). The microsatellite mutation rate used is one that

has been accepted as generally applicable to dinucleotide repeats (which include all our markers) in stickleback and other fishes (Yue *et al.* 2007; Caldera & Bolnick 2008; Berner *et al.* 2009). Although there is considerable uncertainty in the determination of these rates, they have been applied here systematically to all coalescent-based assessments. Consequently, estimates presented are relative to one another, and although not necessarily exact, they still likely reflect relative migration rates among populations. In addition, many coalescent-based estimates have lower high probability density distributions limits (HPD95; Hey 2010) that do not include zero. Final population-specific long-term N_e was calculated by taking the geometric mean of all values determined from pairwise comparisons including the focal population. The proportion of migrants per generation emanating from a focal population was also recovered from all pairwise comparisons (*C* x *V*; see IMa2 manual). We then used all comparisons including a focal population to estimate weighted migration rates to all other populations using the following formula:

$$\mathrm{m}_{i\to j} = \frac{\overline{m_{i\to j}}\; m_{i\to j}}{\sum_{j=1}^{n} m_{i\to j}}, \quad j \neq i \qquad (1)$$

where $m_{i\to j}$ is the per generation migration estimate from population *i* into population *j* determined from the IM model, $\overline{m_{i\to j}}$ is the mean per generation migration rate over all comparisons including population *i*, and *n* is the number of populations considered. The above formula then simplifies to:

$$\mathrm{m_{i\to j}} = \frac{m_{i\to j}}{n}, i \neq j \qquad (2)$$

Although we recognize the simplistic nature of our conversion, which likely fails to consider how migration rates among all populations can interact, it nevertheless makes some concessions

for the uneven distribution of migrants to the different populations and generates per generation migration rates qualitatively comparable to those generated using contemporary methods as implemented in BayesAss 3.0. The advantage of using IM models, however, is that determined parameters are estimated in the coalescent, or over the time frame since populations split (Hey 2010).

Tests of hybrid origin

Because four of the recovered populations within Switzerland corresponded to the MCLs, we tested whether the remaining three populations were of hybrid origin among them. First, the genetic variance-covariance matrix (VCV) of MCL populations, likely representing parental lines, are expected to be less variable and more constrained relative to those of putative hybrid populations (Grant 1994; Steppan *et al.* 2002; Jones *et al.* 2003; Eroukhmanoff & Svensson 2011; Seehausen *et al.* 2014). To test this we performed a principal coordinates analysis (PCoA) in GenalEx on the genetic distances calculated among all individuals. Resulting individual scores along the first two PCo axes were plotted by population in common genotypic space and the area and eccentricity of population-specific 95% confidence ellipses was estimated. The area of the ellipse surrounding a population outlines its genetic variance, while ellipse eccentricity reflects the degree of constraint applied to this variance (Steppan *et al.* 2002; Jones *et al.* 2003; Eroukhmanoff & Svensson 2011; Seehausen *et al.* 2014). High eccentricities (i.e., $\varepsilon \sim 1$) indicate high covariance in genetic signals among loci and thus narrow genetic trajectories, while low eccentricities ($\varepsilon \sim 0$; i.e., a more rounded ellipses) imply less genetic covariance among loci and thus fewer genetic constraints (Steppan *et al.* 2002; Jones *et al.* 2003; Eroukhmanoff & Svensson 2011). PCoAs were also conducted on each Swiss population separately to recover eccentricities

in global genotypic space unconstrained by the variance of other populations. Population-specific ellipse construction and determination of areas and eccentricities were performed in R 3.1.2 (R Core Development Team 2014).

Next, we tested whether the genetic composition of the three putative hybrid populations was of some combination among all MCLs, and whether their admixture proportions was predictable by their spatial arrangement among and/or geographic proximities to MCLs. Alternatively, these populations could trace their ancestries to other lineages outside Switzerland, in which case our predictions would not apply. To test this we simulated an independent hybrid scenario where the genotypes of 50 individuals at 17 loci in 3 populations were generated using EASYPOP 2.0.1 (Balloux 2001). Simulations assumed random mating among diploid individuals with equal proportion of both sexes and where all loci were assumed to evolve at similar rates and following a similar evolutionary model ($\mu = 1 \times 10^{-4}$, combined 85% stepwise mutation, 15% infinite allele models). The number of alleles at each locus was set using levels found in Swiss populations. Simulated populations were connected through a strict island model with relatively low migration rates (0.01 migrants per generation) and allowed to interact for 140 generations. Resulting populations were considered representative of the MCLs and used to generate 3 additional but different hybrid populations (of equal size) using Hybridlab 1.0 (Nielsen *et al.* 2006). The hybrids reflected the anticipated mix among simulated MCLs with the last cross (last population added to the mix) exerting the strongest influence. A list of expected hybrids among simulated MCLs is available (Table S3). Shortest waterway distances (SWD) between each population pairs was also calculated using Google Earth (Google Inc. MountainView CA, USA) measuring distances between the closest sampling locations between populations (see Figs. 1 and

2). In situations where populations were not connected by waterways, shortest overland distances (max < 1 km) between connecting waterways were incorporated in SWD estimates. Both linearized F_{ST} and D_{Jost} estimates of genetic differentiation were compared to log transformed SWDs (to account for multiple dispersal directions and dimensions; Rousset 1997) and to expected genetic differentiation within a hybrid scenario by linear regression analyses supported by 10 000 Mantel randomizations. The combined effects of both SWD and the hybrid scenario were also tested (Revell 2012). Changes to the Akaike information criteria (corrected for small sample sizes; ΔAIC_c) were used to determine the model that best explained genetic differentiation among populations. Mantel regressions were performed in R, where the multivariate versions used the phytools package (Revell 2012).

Finally, we determined whether the genotypes of the putative hybrid populations were consistent with possible combinations of genotypes found in the MCLs, and whether or not they were consistent with a hybrid swarm. We first used all individuals assigned to the MCL populations by STRUCTURAMA/STRUCTURE and tested how successfully they reassigned to their respective populations using exclusion-based assignments in Geneclass2.0 (Rannala & Mountain 1997; Piry *et al.* 2004). Individuals were treated as unknowns and either excluded ($P<0.05$) or considered likely residents of populations using 1 000 000 simulated individuals calculated as per Paetkau *et al.* 2004 (i.e., assuming random mating and based on observed genotypic frequencies within populations). Here, resident/reassignment is defined as the failure to be excluded from a population ($P>0.05$)—that is, an individual cannot be excluded from a population at the 95% level. The successful reassignment of MCL individuals as residents to their respective populations implies that these make good reference populations useful for

excluding individuals of unknown origin (Piry *et al.* 2004; Taylor *et al.* 2006). Next, actual MCL populations were used to generate 50 individuals of various hybrid classes among them including F1s (F1), F1-backcrosses (F1B), F2s (F2), and complex F2s and F2 backcrosses combining all three MCLs (F2C). In all, 17 different hybrid classes were generated from the MCL populations using Hybridlab (Table S4). We then used the MCL populations and the different hybrid classes as reference populations to assign all individuals from the three putative hybrid populations using the same exclusion-based method described above with the same parameters. Individuals that cannot be excluded entirely from various hybrid classes support a hybrid origin of these populations while assignments to complex F2 hybrids and backcrosses is consistent with an origin from within a hybrid swarm combining more than two lineages. We also included individuals collected from the COR and NGG locations as controls to test whether individuals tracing their ancestry outside the MCLs would be excluded from them and their simulated hybrids. To further verify that assignments to more complex hybrid classes are not just an artifact of GeneClass2.0's low power to correctly classify highly admixed individuals, we generated 6 different hybrid classes between Non-Swiss outgroups (COR and NGG) with 50 individuals in each class for a total of 300 individuals. We then used these in the same exclusion-based assignment tests described above to determine whether or not these more complex hybrids would be spuriously assigned to the MCLs or to any one of the 17 different MCL based hybrid classes.

Results

Population genetic structure

The most probable number of genetic populations recovered from unsupervised STRUCTURAMA searches, considering the entire dataset, was six (Table 1, Fig. S1). Using

STRUCTURE to visualize this result showed that most individuals could be assigned to one of these populations with high certainty, with only 5% of individuals assigned to their most probable population with less than 60% probability (32/634) (Fig. 2a). Recovered genetic clusters did not correspond to river drainages, lake systems or sampling sites but rather grouped several sites and certain lake systems, some within different drainages, into the same genetic population (Fig. 2a). One population in particular spanned two different drainages (i.e., the Rhône and the Aare; Orange cluster). Populations at the base of the hierarchy showed some association with colonizing maternal lineages in different areas (Figs. 1 and 2a). Individuals collected from ALL, STS, GLA, GUP, YVB, YVM and WBB showed genetic affiliation with mtDNA lineages found in the Rhône (hereafter Rhône). Individuals collected from MOE, in the upper Rhine, showed genetic affiliation with the purported native Swiss lineage (hereafter MOE), while those collected from GIP, CLA and CUP (hereafter Rhine) showed affiliation with the eastern European lineage present in the lower Rhine (Fig. 2a). Individuals collected from the Lakes Biel/Wohlen region (MOR, GOL, WOH, EYM, GAE, and CHR) formed a genetically distinct population (hereafter WOH; Figs. 1 and 2a). The individuals collected in Corsica and northern Germany also formed genetically distinct populations (hereafter COR and NGG), but we also recognize some level of uncertainty in assignment present among all recovered populations likely reflecting allele sharing due to incomplete lineage sorting and/or admixture (Fig. 2a).

Subsequent STRUCTURAMA analyses performed on all six populations showed variable levels of internal sub-structure. Whereas neither WOH nor COR showed further sub-division, the Rhône, MOE, Rhine, and NGG populations showed additional structure (Table S5). Assignments

of individuals within respective populations as determined in STRUCTURE, largely confirmed STRUCTURAMA results (Fig. 2b-g). In the Rhône population, assignments predominantly grouped individuals collected from Lake Geneva, its tributaries and those at WBB into a population (hereafter RHO) separate from another population (hereafter NEU) made up of individuals mostly collected in Lake Neuchâtel but also present in Lake Geneva and its tributaries (Table S5; Fig. 2b). This likely reflects the higher and more consistent levels of admixture of NEU individuals, with some genetic similarities with individuals in the Rhine and in the distant NGG populations (Fig. 2a-b). More importantly however, this also implies the sympatric coexistence of two genetically distinguishable populations within the Lakes Geneva/ Neuchâtel systems. Additional testing performed on either RHO and NEU revealed no further structure within them. Assignments in the Rhine population separated individuals collected from GIP from those collected in the Lake Constance area (CLA and CUP) (Fig. 2e), likely reflecting the higher admixture levels observed between MOE and GIP (Fig. 2a and e). No further structure was recovered in GIP but additional tests on the Lake Constance area samples recovered two additional populations; one associated with the lake (CLA) and another associated with its upstream tributary (CUP), with substantial admixture between them (Fig. 2e). No further sub-structure was evident in the CUP population but the CLA population exhibited still further structure (Table S5), which was generated from the even split of individual genotypes rather than subdivision among individuals (Fig. 2e). Such results are not indicative of population structure but rather likely indicate the programs inability to distinguish between genotypes at sites with low genetic differentiation (i.e., low F_{ST}; Pritchard *et al.* 2000; Falush *et al.* 2007; Hubisz *et al.* 2009). Similarly, although STRUCTURAMA indicated substantial internal genetic structure in MOE and NGG populations (Table S5), more detailed individual assignments tests showed both

cases were examples of genotype splitting (Fig. 2d and f). The overall hierarchical search for population structure therefore, recovered nine genetically distinguishable populations among the 634 sampled individuals. Of these, two were outside of Switzerland (COR and NGG), four were consistent with the main colonizing lineages (RHO, MOE and CLA-CUP), and the last three (NEU, WOH and GIP), although genetically distinguishable by microsatellite allele frequencies, exhibited various mtDNA haplotypes (Figs. 1 and 2).

Neutrality tests

None of the markers used to recover population genetic structure at the different hierarchical levels showed evidence of selection using the FDIST algorithm as applied in LOSITAN, regardless of the applied mutational model (Fig. S2). This includes all seven microsatellite markers linked to quantitative traits in other populations (Peichel *et al.* 2001; Mäkinen *et al.* 2008a). Similarly, selection tests using Bayescan 2.1 also failed to detect signs of selection in any used markers (Fig. S2). These results indicate that neutral processes largely governed allele frequencies and population genetic differentiation at the markers used.

Population genetic statistics

Descriptive statistics of genetic diversity over the nine populations and 17 loci are available (Table S6). No evidence of linkage disequilibrium was detected between any pair of loci ($p >$ 0.05). Eight population-loci combinations deviated from genotypic frequencies expected under HWE, out of a possible 153 comparisons, a number very close to that expected by chance ($n =$ 7.65). None of these deviations involved the same locus in different populations consistent with their random nature. The 17 loci showed variable levels of polymorphism in the different

populations. The allelic richness (A_R) ranged between 1.00 and 9.80 with a mean of 3.24, and the number of private alleles (*Pa*) ranged from 0.00 to 1.29 with a mean of 0.38, over all populations and loci. Large and significant levels of genetic differentiation estimated as F_{ST} and D_{Jost} were detected among all possible pairwise population comparisons, indicating strong support for genetic differences among them (Table S7). These differences were generally greater among populations reflecting the MCLs. No significant differences were found in population genetic diversity indices or global F_{ST} estimated using putatively QTL linked versus unlinked loci ($W \geq 25$, $p \geq 0.216$), consistent with marker neutrality. No significant differences were observed in genetic diversity indices among the MCL populations versus those exhibiting mixed mitochondrial lineages ($t \leq 2.00$, *d.f.* range = 3.01-4.95, $p \geq 0.164$).

Population sizes and connectivity

All nine recovered populations exhibited comparable contemporary N_e except WOH and COR, which had estimates near an order of magnitude greater (Fig. 3). The WOH population was by far the largest within Switzerland while CLA was the smallest. These results were similar when considering a greater minimum allele frequency of 0.02, except that the estimates for COR became indeterminate (Fig. S3). Only three populations were connected by contemporary migration rates greater than 0.01 (Fig. 3). These higher migration rates showed high unidirectional migration from CUP to CLA and more restricted unidirectional migration from CUP to GIP. Thus, CUP acts as a source population to both GIP and CLA. All other populations appear contemporarily isolated. To eliminate the possibility that low contemporary migration rates are an artifact of the way we grouped individuals into populations (i.e., by assignment probability), we also estimated migration rates using individuals grouped by sample location.

Here, individuals were assigned to populations based on the predominant genetic cluster recovered at each site. Contemporary migrations rates produced in this way were nearly identical except that we also recovered some low migration (0.014) from RHO into NEU (see Fig S4).

Coalescent-based N_e estimates tended to be smaller and less variable than contemporary ones ($\sigma_{contemporary}$ = 490.5, $\sigma_{coalescent}$ = 137.4) and showed that most populations were of comparable size (Fig. 4). Unlike estimates of contemporary gene flow, coalescent-based per generation migration rates showed extensive (>> 0.01) multidirectional gene flow among populations within Switzerland (Fig. 4). Notably, most Swiss populations consistent with MCLs (i.e., RHO, MOE, CLA and CUP) tended to export more and import fewer migrants than did the populations of putative hybrid origins (NEU, WOH, GIP). We found no indications of historical gene flow between any Swiss population and the Corsican one, and the possibility of low historical gene flow between a single Swiss population (RHO) and the North German one. This may be a legacy of gene flow in the original range of these populations outside Switzerland.

Tests of hybrid origin

PCo analyses performed on the genetic distances among individuals collected within Switzerland showed distinct clustering of individuals belonging to the seven Swiss populations with variable degrees of overlap (Fig. 5). MCL populations tended to occupy the periphery of the genotypic space outlined by the first 2 PCo axes (accounting for nearly 70% of the genetic variation among individuals), while the remaining three populations (NEU, WOH, GIP) were encompassed entirely within the range defined by the MCL populations. The area of the 95% confidence ellipses calculated for the MCL populations were significantly smaller than those calculated for

the remaining three consistent with greater genetic variation in the latter group and with their hybrid origin (t = 3.391, $d.f.$ = 4.16, p = 0.013). The ellipses of the three remaining populations were also less eccentric relative to those of the MCLs when compared both in common (t = 3.883, $d.f.$ = 2.03, p = 0.029) and global (t = 2.231, $d.f.$ = 4.01, p = 0.047) genotypic spaces, consistent with relaxed genetic constraints and increased evolutionary potential expected in hybrids. Differences in ellipse areas and eccentricities assessed in common genotypic space were still significant after adjusting p-values for multiple comparison using the Benjamini-Yekutieli (2001) correction which accounts for false discoveries among rejected hypotheses (Benjamini & Yekutieli 2001; p = 0.0390 and p = 0.0435 for areas and eccentricities, respectively).

Results of the AIC_c model comparisons of F_{ST} and D_{JOST} based Mantel regressions showed similar results (Table 2). In both cases, the most likely model explaining genetic differentiation among Swiss populations was one based solely on the hybrid scenario, while that using shortest waterway distances exclusively, or in combination with the hybrid scenario were less likely and/or not significant (Table 2). These results imply the uneven and variable contribution of the different MCLs to the various possible hybrid populations, and that this contribution is more likely related to the spatial arrangement of the MCLs within Switzerland, rather than to the strict distances between them.

Nearly 90% of individuals from each MCL population could not be excluded from their respective population at the 0.05 level (Fig. 6). In all cases, only exclusion errors were made and no individual was incorrectly reassigned to one of the other MCL populations, indicating that the MCLs were suitable reference populations for exclusion analyses of unknown individuals (Fig.

6). Using the MCLs and simulated hybrid classes in exclusion analyses performed on individuals tracing their ancestry in populations located outside Switzerland (COR and NGG) and their hybrids showed that all individuals were excluded from both the MCLs and their expected hybrid classes (Figs. 6d, e and S5). Consequently, assignment of actual purported hybrid individuals to expected hybrid classes is not likely an artifact of GeneClass2.0's limited power to assign/exclude highly admixed individuals. Performing the same analyses on NEU individuals showed that over 25% could not be excluded from the RHO population (Fig. 6f). This result is not surprising given the similarity between RHO and NEU (see Figs. 2 and 5). Moreover, a substantial proportion of NEU individuals could also not be excluded from possible hybrid classes with a general increase in assignment probabilities as the hybrid class complexity increased (Fig. 6f). Similar exclusion tests performed on WOH and GIP showed that all individuals were excluded from all MCL populations (Fig. 6g and h). On the other hand, a substantial proportion of both WOH and GIP individuals could not be excluded from possible hybrid classes, and the same general pattern of increasing assignment probabilities with increasing hybrid complexity was observed.

Discussion

Here, we show that the recent range expansion of threespine stickleback in Switzerland is associated with the formation of a hybrid 'superswarm' among three distinct lineages that colonized Switzerland about 140 years ago (Heller 1870; Fatio 1882; Bertin 1925; Lucek *et al.* 2010). This massive hybridization likely gave rise to three genetically distinguishable novel populations. We demonstrate that current populations are genetically stable and all but the most closely related ones seem nearly isolated with low levels of contemporary gene flow. Coalescent-

based analyses on the same populations, however, show clear connectivity with extensive multidirectional gene flow among them in the recent past. If our inferences are correct, backcrossing to the source populations is less than expected from geographical distances, and migration between areas that currently host genetically differentiated populations of hybrid origin seems lower now than during colonization. Thus, it appears as though secondary contact among three distant lineages during the colonization of Swiss waterways initially led to formation of a hybrid 'superswarm', followed by stabilization of genetically differentiated populations. Whether or not this hybridization among the main colonizing lineages and the stabilization of hybrid populations has facilitated ecological range expansion into various habitats remains to be determined (Lucek et al. 2010; Lucek et al. 2014)

Population Structure

Population Structure analyses (STRUCTURAMA/STRUCTURE) identified seven genetic stickleback populations from our hierarchical analyses of samples taken from 17 sites within Switzerland. The approach we used to infer population structure differs from many previous population based stickleback studies, some performed in these systems (Reusch *et al.* 2001; Mäkinen *et al.* 2006; Lucek *et al.* 2010; Lucek *et al.* 2013; Lucek *et al.* 2014a; Lucek *et al.* 2014b). Rather than assigning population status to different sampling sites by default, we used an approach based on individual admixture proportions. Although both methods are effective, they are useful in addressing different hypotheses. In the context of reconstructing a biological invasion from multiple introductions of distantly related lineages, an approach using an individual-based population genetics framework (i.e., individuals assigned to population in HWE with low linkage among loci) may be more appropriate (Darling *et al.* 2008).

The recovered population structure groups several geographically distant locations together within the same genetic population, irrespective of habitat type (Fig 2). The RHO population in particular, spanned more than a single drainage and had a disjunct distribution, where pockets of individuals were isolated from one another by regions occupied by the NEU or WHO populations (i.e., RHO individuals at the WBB sampling site). This disjunct distribution is likely the result of translocation and subsequent isolation of RHO individuals in the WBB area. The general lack of site-specific population structure indicates substantially greater gene flow among sampling locations and habitat types within recovered genetic populations relative to that among them. On the other hand, our analysis also occasionally assigns individuals within single sampling sites to two different genetic populations, suggesting perhaps that distinct stickleback genetic populations coexist at some sites in the Lakes Neuchâtel and Geneva systems.

The population genetic structure recovered here, as inferred by the analyses described above, cannot be readily explained by local adaptation to distinct habitats but rather likely reflect the processes of colonization and gene flow. First, two outlier loci detection approaches (LOSITAN-FDIST and Bayescan) found no evidence of diversifying or balancing selection at any loci. Second, even though some of our markers were shown to be linked to known QTLs in studies of other stickleback populations (Peichel *et al.* 2001; Mäkinen *et al.* 2008a), these loci did not behave differently from neutral markers. This result supports the notion that quantitative traits linked to specific markers determined in some populations does not necessarily imply these same marker will show similar trait-based relationships in other populations (Peichel *et al.* 2001; Mäkinen *et al.* 2008a).

Population Connectivity and Size

Extensive contemporary gene flow among populations would likely result in violations of HWE and/or LD among loci within populations greater than expected by chance alone (e.g., heterozygote deficiencies). This could result in Wahlund effects within populations or in signs of recombination or epistatic linkage among loci (Slatkin 2008; Excoffier & Lischer 2010). Without exception, however, no departures from HWE or evidence of excessive LD are evident in our recovered genetic populations. Moreover, our genetic populations are significantly differentiated, often showing high F_{ST}/D_{Jost} indices, with no indication of contemporary gene flow among them. The only contemporary gene flow observed here occurs in a unidirectional manner from CUP into both CLA and GIP. These results are in accordance with previous work showing substantial gene flow among stickleback collected from stream and lake locations within the Lake Constance region (Berner *et al.* 2010; Moser *et al.* 2012; Lucek *et al.* 2013; Lucek *et al.* 2014b) and between Constance region stickleback and those in the upper Rhine (i.e., GIP; Lucek *et al.* 2010). Lucek *et al.* 2014b) suggest that stickleback within the Constance region have become divergently adapted with decreasing gene flow between lake and stream populations. So, gene flow observed between CLA and CUP is likely occurring in primary contact between diverging stream and lake ecotypes that originated within the past 140 years from a common gene pool. Coalescent-based analyses support the gene flow reduction in the Constance region in particular, but also more generally throughout Switzerland. IM based coalescent analyses suggest extensive multidirectional gene flow among most Swiss populations and recovers much larger migration estimates than methods used to estimate contemporary gene flow. The differences between estimated per generation migration rates are likely due to methods for assessing contemporary gene flow only taking current allele frequencies into account and thus only resolving recent

migration among populations (Wilson & Rannala 2003; Piry *et al.* 2004). Coalescent-based analyses as implemented in IMa2, instead, estimate migration rates over the divergence time between and among populations (Hey & Nielsen 2004; Hey 2010; Strasburg & Rieseberg 2010). The latter are essentially averages over the coalescent and do not make concessions for migration rates that may be temporally dynamic. Thus, coalescent-based migration rate estimates can be quite different from those using contemporary methods, which reflect more current population connectivity. Here, we combined both approaches, which together suggest that although gene flow among recovered Swiss genetic populations was extensive in the past, it has likely been substantially reduced relatively recently. Coalescent-based estimates show that populations corresponding to the three main colonizing lineages (RHO, MOE, CLA/CUP) exhibit much larger outgoing than incoming migration rates while the opposite pattern holds for the remaining three populations (NEU, WOH and GIP). Consistent with previous findings (Lucek *et al.* 2010), our results suggest that the three main colonizing lineages, geographically restricted to the northeast, northwest and far west parts of Switzerland, acted as genetic sources seeding other populations as they expanded across the Swiss midlands and now show variable levels of complex admixture among main colonizing lineages.

Hybrid superswarm

Given the high level of gene flow that the putative hybrid populations (NEU, WOH and GIP) received from the MCLs in the past, a plausible scenario for their origin is genetic admixture among the MCLs. As expected, these populations occupy intermediate and less constrained (more variable) genotypic space than the MCLs, consistent with the breakdown and reshuffling of genetic constraints established in parental lineages (Buerkle *et al.* 2000; Mallet 2007; Schluter

& Conte 2009; Abbott *et al.* 2013; Seehausen *et al.* 2014). Assignment tests also showed improving assignments of individuals in hybrid populations to increasingly complex simulated hybrid classes. Exclusion-based assignments allow individuals to remain unclassified if their genotype is too dissimilar from the reference populations (Paetkau *et al.* 2004; Piry *et al.* 2004). Consequently, finding an increasing number of individuals assigned to increasingly complex hybrid classes implicates admixture among all three MCLs in the formation of these three populations. It is important to note that while assignment to hybrid classes may be relatively low, we tested only 17 of a diverse array of hybrid classes potentially produced by the MCLs and included only formative F1s and F2s and their backcrosses. Consequently, tests including more complex hybrid classes may find greater hybrid assignment. Moreover, relatively low assignment rates may also reflect past hybridization followed by genetic stabilization and recombination within newly established hybrid populations possibly eroding more obvious hybridization signals (Currat *et al.* 2008; Seehausen *et al.* 2008). This is supported by the NEU population, which is the least differentiated among the hybrid populations showing the highest hybrid assignments. This may indicate that, all else being equal, and in light of the limited contemporary gene flow (see above), the NEU population is the most recently formed hybrid. On the other hand, NEU is also the only hybrid population para- or sympatrically distributed in some sites with the RHO MCL population. Consequently, its greater assignments to hybrid classes may also be related to its continued physical contact and possible gene flow with, a seeding colonizing lineage (see Fig. S4; Lucek *et al.* 2010; Lucek *et al.* 2014a), whereas GIP and WOH are currently entirely allopatric from all MCLs as determined here.

The hybrid origin of NEU, WOH and GIP is also consistent with modeling results showing the best model explaining genetic differentiation among populations is one explicitly assigning intermediate genetic makeup to putative hybrid populations relative to simulated MCLs. We also found no relationship between genetic differentiation and geographical distance either in combination with the hybrid scenario or by itself. Contrary to previous work performed within Swiss lake systems (Lucek *et al.* 2013), the pattern of genetic population divergence observed here is not likely driven by habitat dependent selection because, although genetically differentiated, many populations occupy similar (e.g., GIP-MOE, both stream habitats), or even the same, habitats (e.g., both RHO- and NEU-like individuals recovered from STS, ALL, GLA, GUP, YVB and YVM). Thus, although parallel habitat based divergence seems evident at a finer, more lake-specific level (Lucek *et al.* 2013), the nature of genetic boundaries between the geographically more inclusive genetic populations identified in the present work is less obvious. Clarifying the causes of among population divergence and the apparent reduction/cessation of gene flow among them as recovered here, as well as testing whether some populations do in fact locally coexist as distinct genetic clusters, are logical next steps for future work.

Irrespective of the mechanisms, the hybrid origins of NEU, WOH and GIP populations is consistent with previous reports implicating differential hybridization as an important cause of among-population divergence in genotype and phenotype (Lucek *et al.* 2010) and as facilitating invasions more generally (Lockwood *et al.* 2007; Prentis *et al.* 2008; Lack *et al.* 2012; Parepa *et al.* 2014; Williams *et al.* 2014). An important distinction from many previous reports, however, is that we show evidence of three genetically-distinguishable populations originating from a

hybrid 'superswarm' involving complex crosses and backcrosses among more than two distant lineages and differing in their degree of genetic differentiation.

Conclusion

Our findings supports the formation of stickleback hybrid populations that have contributed to the extensive genetic and likely phenotypic (Lucek *et al.* 2010; Lucek *et al.* 2013) diversity observed on small spatial scales within Switzerland. This is consistent with secondary contact among distant lineages converting interpopulation genetic diversity into intrapopulation genetic variation by hybridization (Lockwood *et al.* 2007; Dlugosch & Parker 2008; Prentis *et al.* 2008; Seehausen *et al.* 2008). We show that this process occurred among three lineages, but to varying degrees in different places and likely provided extensive standing genetic variation that facilitated range expansion. Here, three new populations of hybridogenic origins have likely emerged in Switzerland over the last 140 years of secondary contact. Our study adds to a growing body of work implicating hybridization as a facilitator of range expansion and possibly the rapid onset of adaptive diversification (Mallet 2007; Nolte & Tautz 2010; Abbott *et al.* 2013; Seehausen *et al.* 2014) over short time scales.

Acknowledgments

Sampling help was provided by the Fish Ecology group at the EAWAG, and in particular by Pascal Vonlanthen, Guy Périat, Alan Hudson and Isabel Magalhaes. We thank the Swiss Cantonal authorities of Aargau, Basel Land, Bern, St.Gallen, Thurgau, Valais and Vaud as well as the French fishery authorities ONEMA (Corsica) for collection permits. The EAWAG Action

Field Grant 'AquaDiverse' and the Swiss National Science Foundation Grant 31003A-118293 to OS supported this work.

Data accessibility

The raw genotypes for all individuals, input files for STRUCTURE 2.3.4/STRUCTURAMA 2.0, input files for both LOSITAN and BayeScan 2.1 along with all IMa2 input files used in this study are stored and accessible through the Dryad data repository: doi:10.5061/dryad.c2n5j

References

Abbott R, Albach D, Ansell S*, et al.* (2013) Hybridization and speciation. *J. Evol. Biol.* **26**, 229-246.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics* **9**, 1-5.

Balloux F (2001) EASYPOP (Version 1.7): A Computer Program for Population Genetics Simulations. *J. Heredity* **92**, 301-302.

Bell MA, Foster SA (1994) *The Evolutionary Biology of the Threespinned Stickleback* Oxford University Press, Oxford, UK.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Berner D, Grandchamp AC, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: Sticklebacks across eight lake-stream transitions. *Evolution* **63**, 1740-1753.

Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Mol. Ecol.* **19**, 4963-4978.

Bertin L (1925) Recherches bionomiques, biométriques et systématiques sur les épinoches (*Gastérostéidés*). *Ann. Inst. Océano.* **II**, 205.

Bradbury IR, Hubert S, Higgins B*, et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evol Appl* **6**, 450-461.

Buerkle CA, Morris RJ, Asmussen MA, Rieseberg LH (2000) The likelihood of homoploid hybrid speciation. *Heredity* **84**, 441-451.

Caldera EJ, Bolnick DI (2008) Effects of colonization history and landscape structure on genetic variation within and among threespine stickleback (*Gasterosteus aculeatus*) populations in a single watershed. *Evol. Ecol. Res.* **10**, 575-598.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: Massive introgression by local genes. *Evolution* **62**, 1908-1920.

Darling JA, Bagley MJ, Roman J, Tepolt CK, Geller JB (2008) Genetic patterns across multiple introductions of the globally invasive crab genus Carcinus. *Mol. Ecol.* **17**, 4992-5007.

Deagle BE, Jones FC, Chan YF*, et al.* (2012) Population genomics of parallel phenotypic evolution in stickleback across stream–lake ecological transitions. *Proc. R. Soc. B-Biol. Sci.* **279**, 1277-1286.

Dieringer D, Schlötterer C (2003) Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167-169.

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* **17**, 431-449.

Do C, Waples RS, Peel D*, et al.* (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol. Ecol. Res.* **14**, 209-214.

Eroukhmanoff F, Svensson EI (2011) Evolution and stability of the G-matrix during the colonization of a novel environment. *J. Evol. Biol.* **24**, 1363-1373.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 564-567.

Falush D, Stephens M, Pritchard J (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574-578.

Fatio V (1882) *Faune des vertébrés de la Suisse*, 1st edn. H. Georg, Genève.

Foll M, Gaggiotti O (2008) A Genome-Scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**, 977-993.

Gerlach G, Jueterbock A, Kraemer P, Deppermann J, Harmand P (2010) Calculations of population differentiation based on GST and D: forget GST but not all of statistics! *Mol. Ecol.* **19**, 3845-3852.

Goudet J (1995) FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. *J. Heredity* **86**, 485-486.

Grant PR (1994) Population Variation and Hybridization - Comparison of Finches from 2 Archipelagos. *Evolutionary Ecology* **8**, 598-617.

Hare M, Nunney L, Schwartz MK*, et al.* (2011) Understanding and estimating effective population size for practical application in marine species management. *Conserv. Biol.*

Heller C (1870) Die Fishes Tirols und Vorarlbergs. *Z. Ferdinandeums Tirol* **5**, 295-369.

Hey J (2010) Isolation with Migration Models for More Than Two Populations. *Mol. Biol. Evol.* **27**, 905-920.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D-persimilis. *Genetics* **167**, 747-760.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Res.* **9**, 1322-1332.

Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. *Genetics* **175**, 1787-1802.

Huelsenbeck JP, Andolfatto P, Huelsenbeck ET (2011) Structurama: Bayesian inference of population structure. *Evol. Bioinf.* **7**, 55.

Jakobsdóttir KB, Pardoe H, Magnússon Á*, et al.* (2011) Historical changes in genotypic frequencies at the Pantophysin locus in Atlantic cod (*Gadus morhua)* in Icelandic waters: evidence of fisheries-induced selection? *Evol. Appl.* **4**, 562-573.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jones AG, Arnold SJ, Borger R (2003) Stability of the G-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. *Evolution* **57**, 1747-1760.

Jones FC, Grabherr MG, Chan YF*, et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61.

Kaeuffer R, Peichel CL, Bolnick DI, Hendry AP (2012) Parallel and Nonparallel Aspects of Ecological, Phenotypic, and Genetic Divergence across Replicate Population Pairs of Lake and Stream Stickleback. *Evolution* **66**, 402-418.

Lack JB, Greene DU, Conroy CJ*, et al.* (2012) Invasion facilitates hybridization with introgression in the Rattus rattus species complex. *Mol. Ecol.* **21**, 3545-3561.

Lockwood JL, Hoopes M, Marchetti MP (2007) *Invasion Ecology* Blackwell Publishing.

Lucek K, Lemoine M, Seehausen O (2014a) Contemporary ecotypic divergence during a recent range expansion was facilitated by adaptive introgression. *J Evol Biol* **27**, 2233-2248.

Lucek K, Roy D, Bezault E, Sivasundar A, Seehausen O (2010) Hybridization between distant lineages increases adaptive variation during a biological invasion: stickleback in Switzerland. *Mol. Ecol.* **19**, 3995-4011.

Lucek K, Sivasundar A, Roy D, Seehausen O (2013) Repeated and predictable patterns of ecotypic differentiation during a biological invasion: lake–stream divergence in parapatric Swiss stickleback. *J Evol Biol* **26**, 2691-2709.

Lucek K, Sivasundar A, Seehausen O (2014b) Disentagling the role of phenotypic plasticity and genetic divergence in contemporary ecotype formation during a biological invasion. *Evolution* **68**, 2619-2632.

Mäkinen HS, Cano JM, Merilä J (2006) Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. *Mol. Ecol.* **15**, 1519-1534.

Mäkinen HS, Cano M, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Mol. Ecol.* **17**, 3565-3582.

Mäkinen HS, Merilä J (2008) Mitochondrial DNA phylogeography of the three-spined stickleback (Gasterosteus aculeatus) in Europe - Evidence for multiple glacial refugia. *Mol. Phylogenet. Evol.* **46**, 167-182.

Mallet J (2007) Hybrid speciation. *Nature* **446**, 279-283.

McPhail JD (1984) Ecology and evolution of sympatric sticklebacks (Gasterosteus): morphological and genetic evidence for a species pair in Enos Lake, British Columbia. *Can. J. Zool.* **62**, 1402-1408.

Moser D, Roesti M, Berner D (2012) Repeated Lake-Stream Divergence in Stickleback Life History within a Central European Lake Basin. *PLoS ONE* **7**, e50620.

Munzing J (1963) Evolution of variation and distributional patterns in European populations of 3-spined stickleback, Gasterosteus aculeatus. *Evolution* **17**, 320-332.

Nielsen EE, Bach LA, Kotlicki P (2006) HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes* **6**, 971-973.

Nolte AW, Tautz D (2010) Understanding the onset of hybrid speciation. *Trends in Genetics* **26**, 54-58.

Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol. Ecol.* **13**, 55-65.
Parepa M, Fischer M, Krebs C, Bossdorf O (2014) Hybridization increases invasive knotweed success. *Evol. Appl.* **7**, 413-420.
Peakall ROD, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6**, 288-295.
Peichel CL, Nereng KS, Ohgi KA*, et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901-905.
Piry S, Alapetite A, Cornuet JM*, et al.* (2004) GENECLASS2: A software for genetic assignment and first-generation migrant detection. *J. Heredity* **95**, 536-539.
Prentis PJ, Wilson JRU, Dormontt EE, Richardson DM, Lowe AJ (2008) Adaptive evolution in invasive species. *Trends in Plant Science* **13**, 288-294.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945-959.
Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *P Natl Acad Sci USA* **94**, 9197-9201.
Ravinet M, Prodohl PA, Harrod C (2013) On Irish stickleback: morphological diversification in a secondary contact zone. *Evol. Ecol. Res.* **15**, 271-294.
Reusch TBH, Wegner KM, Kalbe M (2001) Rapid genetic divergence in postglacial populations of threespine stickleback (*Gasterosteus aculeatus*): the role of habitat type, drainage and geographical proximity. *Mol. Ecol.* **10**, 2435-2445.
Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217-223.
Rice WR (1989) Analyzing Tables of Statistical Tests. *Evolution* **43**, 223-225.
Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137-138.
Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228.
Roy D, Hardie DC, Treble MA, Reist JD, Ruzzante DE (2014) Evidence supporting panmixia in Greenland halibut (*Reinhardtius hippoglossoides*) in the Northwest Atlantic. *Can. J. Fish. Aquat. Sci.* **71**, 763-774.
Sakai AK, Allendorf FW, Holt JS*, et al.* (2001) The population biology of invasive species. *Annu. Rev. Ecol. Evol. Syst.* **32**, 305-332.
Schluter D (1993) Adaptive radiation in sticklebacks - size, shape, and habitat use efficiency. *Ecology* **74**, 699-709.
Schluter D (1996) Adaptive radiation along genetic lines of least resistance. *Evolution* **50**, 1766-1774.
Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proc. Natl. Acad. Sci. USA* **106**, 9955-9962.
Seehausen O (2004) Hybridization and adaptive radiation. *Trends Ecol Evol* **19**, 198-207.
Seehausen O, Butlin RK, Keller I*, et al.* (2014) Genomics and the origin of species. *Nat Rev Genet* **15**, 176-192.
Seehausen O, Takimoto G, Roy D, Jokela J (2008) Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Mol. Ecol.* **17**, 30-44.

Simberloff D (2009) The Role of Propagule Pressure in Biological Invasions. *Ann. Rev. Ecol. Syst.* **40**, 81-102.
Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477-485.
Steppan SJ, Phillips PC, Houle D (2002) Comparative quantitative genetics: evolution of the G matrix. *Trends. Ecol. Evol.* **17**, 320-327.
Strasburg JL, Rieseberg LH (2010) How robust are "isolation with migration" analyses to violations of the im model? A simulation study. *Mol Biol Evol* **27**, 297-310.
Taylor E, Boughman J, Groenenboom M*, et al.* (2006) Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. *Mol. Ecol.* **15**, 343-355.
Thompson CE, Taylor EB, McPhail JD (1997) Parallel evolution of lake-stream pairs of threespine sticklebacks (Gasterosteus) inferred from mitochondrial dna variation. *Evolution* **51**, 1955-1965.
van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**, 535-538.
Waples RS, England PR (2011) Estimating Contemporary Effective Population Size on the Basis of Linkage Disequilibrium in the Face of Migration. *Genetics* **189**, 633-644.
Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358-1370.
Williams WI, Friedman JM, Gaskin JF, Norton AP (2014) Hybridization of an invasive shrub affects tolerance and resistance to defoliation by a biological control agent. *Evol. Appl.* **7**, 381-393.
Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177-1191.
Yoder JB, Clancey E, Des Roches S*, et al.* (2010) Ecological opportunity and the origin of adaptive radiations. *J. Evol. Biol.* **23**, 1581-1596.
Yue G, David L, Orban L (2007) Mutation rate and pattern of microsatellites in common carp (*Cyprinus carpio* L.). *Genetica* **129**, 329-331.

Table 1. Population structure estimated in sampled stickleback determined from unsupervised searches (performed in STRUCTURAMA 2.0). *EK* values indicate Dirichlet Process Prior mean on which searches were centered. Marginal likelihood of searches indicates the likelihood of the resulting search performed using the corresponding *EK*.

K	*EK(1)*	*EK(2)*	*EK(3)*	*EK(4)*	*EK(5)*	*EK(6)*	*EK(7)*	*EK(8)*	*EK(9)*	*EK(10)*	*EK(12)*
Over all sampled sites											
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	**0.67**	**0.67**	0.33	0.33	0.00	0.33	0.33	0.00	0.00	0.00	0.00
4		0.33	0.00	**0.34**	0.33	0.00	**0.34**	0.33	0.00	0.33	0.00
5			**0.67**	0.33	**0.67**	**0.67**	**0.34**	**0.67**	0.00	**0.34**	0.33
6									**1.00**	0.33	**0.67**
7											
	Marginal likelihood of search										
	-25898.19	-24585.50	-24584.81	-24584.19	-20386.27	-24584.80	-24584.79	-20386.95	**-17734.47**	-20385.77	-18819.18

Most likely number of recovered clusters is bolded
Bolded marginal likelihood of searches indicate most robust and likely search results

Table 1. Concluded.

K	*(EK15)*	*EK(17)*	*EK(20)*
Over all sampled sites			
1	0.00	0.00	0.00
2	0.00	0.00	0.00
3	0.00	0.00	0.00
4	0.00	0.00	0.00
5	**0.67**	0.00	0.33
6	0.33	**1.00**	**0.67**
7			0.00
	Marginal likelihood of search		
	-19315.31	**-17734.67**	-18820.36

Table 2. Regression models explaining the genetic differentiation among Swiss stickleback populations. n = number of populations in the model, K = number of explanatory variables, R^2 = coefficient of determination, P_{ols} = ordinary least squared *P* value, P_m = Mantel permutations *P* values (10 000), and RSS = residual sum of squares. Variables in the models are: ln(SWD) = log transformed shortest waterway distance between populations and *Hybsc* = matrix of expected genetic differences under the hybrid scenario considering exNEU, exWOH and exGIP as hybrid populations originating from crosses among simulated main colonizing lineages (sRHO, sMOE, and sCON). Most likely models are bolded.

Model	n	K	R^2	P_{ols}	P_m	RSS	AIC_c	ΔAIC_c
F_{ST} ~ ln(SWD)	6	1	0.153	0.149	0.072	0.629	-10.534	2.025
F_{ST} ~ *Hybsc*	**6**	**1**	**0.396**	**0.012**	**0.050**	**0.449**	**-12.558**	**0.000**
F_{ST} ~ ln(SWD) + *Hybsc*	6	2	0.400	0.047	0.086	0.446	-7.597	4.963
D_{Jost} ~ ln(SWD)	6	1	0.279	0.043	0.035	3.134	-0.897	2.324
D_{Jost} ~ *Hybsc*	**6**	**1**	**0.511**	**0.003**	**0.013**	**2.127**	**-3.221**	**0.000**
D_{Jost} ~ ln(SWD) + *Hybsc*	6	2	0.537	0.010	0.015	2.015	1.453	4.674

Figure 1. Detailed view of 17 locations within Switzerland where stickleback were sampled. Main river drainages are coloured (orange = Rhône, blue = *Aare* and green = *Rhine*) and five lake systems (Geneva, Neuchâtel, Wohlen (*not shown*), Biel and Constance). Each site code corresponds to that listed in Table S1 and shows the proportion of mtDNA haplotypes determined in Lucek et al. (2010). CHR was not assessed for mtDNA. Inset map shows Switzerland's location within mainland Europe and the location of the Corsican (COR) and the North German (NGG) sampling sites.

Figure 2. Hierarchical Bayesian posterior probability assignment of sampled stickleback. (a) Initial analysis using all individuals recovered 6 genetic clusters. Subsequent analysis run on recovered clusters (b-g), shows up to 9 genetically distinguishable clusters present in sampled data (7 within Switzerland proper). Each individual is represented by a bar whose colour corresponds to its probability of belonging to recovered genetic clusters. Locations where all genotypes are split indicate all individuals are genetically similar but admixed from multiple sources. Black and white horizontal bars above structure plots delimit main river drainage and lake systems.

Figure 3. Contemporary effective population sizes ($\hat{N}_e$) and migrations rates (m) among recovered populations. Circles represent the ln($\hat{N}_e$)*10 and the shading outlines their upper 95% confidence limit determined from Jackknifing over loci pairs and using allele frequencies greater than 0.01. Contemporary migration rates (m) ≥ 0.01 (i.e., ≥ 1%) are also shown which were determined using BayesAss3.0.

Figure 4. N_e and m estimates determined from coalescent-based analyses performed in IMa2. Circles represent the ln(N_e)*10 and the shading outlines upper high probability density interval similar to 95% confidence limits for Bayesian parameter estimates (HPD95). m rates determined from multiple pairwise comparisons between populations as described in text.

Figure 5. Principal coordinates analyses of genetic distances among sampled Swiss stickleback. Ellipses encircle 95% of the individuals assigned to each genetic population as determined using STRUCTURAMA/STRUCTURE. Numbers in parentheses indicate the amount of variation determined along each axes.

Figure 6. Relative assignment probabilities of sampled stickleback to various potential source populations. Panels a-c show the reassignments of individuals from the RHO, MOE and CLA/CUP populations respectively, representing the main colonizing lineages (MCL). Panels d-h show the assignment of the control NGG and COR, and the tested NEU, WHO and GIP individuals to the main lineages and the various hybrid forms expected between them. F1 = hybrid between two main lineages, F1B = back cross between an F1 hybrid and a main lineage, F2 = the combination of two similar type hybrids and F2C = the combination of two different types of hybrids and backcrosses combining the 3 MCLs.